openai.com

Governance of superintelligence

Authors

5-6 minutes

Given the picture as we see it now, it’s conceivable that within the next ten years, AI systems will exceed expert skill level in most domains, and carry out as much productive activity as one of today’s largest corporations.

In terms of both potential upsides and downsides, superintelligence will be more powerful than other technologies humanity has had to contend with in the past. We can have a dramatically more prosperous future; but we have to manage risk to get there. Given the possibility of existential risk, we can’t just be reactive. Nuclear energy is a commonly used historical example of a technology with this property; synthetic biology is another example.

We must mitigate the risks of today’s AI technology too, but superintelligence will require special treatment and coordination.

A starting point

There are many ideas that matter for us to have a good chance at successfully navigating this development; here we lay out our initial thinking on three of them.

First, we need some degree of coordination among the leading development efforts to ensure that the development of superintelligence occurs in a manner that allows us to both maintain safety and help smooth integration of these systems with society. There are many ways this could be implemented; major governments around the world could set up a project that many current efforts become part of, or we could collectively agree (with the backing power of a new organization like the one suggested below) that the rate of growth in AI capability at the frontier is limited to a certain rate per year.

And of course, individual companies should be held to an extremely high standard of acting responsibly.

Second, we are likely to eventually need something like an IAEA for superintelligence efforts; any effort above a certain capability (or resources like compute) threshold will need to be subject to an international authority that can inspect systems, require audits, test for compliance with safety standards, place restrictions on degrees of deployment and levels of security, etc. Tracking compute and energy usage could go a long way, and give us some hope this idea could actually be implementable. As a first step, companies could voluntarily agree to begin implementing elements of what such an agency might one day require, and as a second, individual countries could implement it. It would be important that such an agency focus on reducing existential risk and not issues that should be left to individual countries, such as defining what an AI should be allowed to say.

Third, we need the technical capability to make a superintelligence safe. This is an open research question that we and others are putting a lot of effort into.

What’s not in scope

We think it’s important to allow companies and open-source projects to develop models below a significant capability threshold, without the kind of regulation we describe here (including burdensome mechanisms like licenses or audits).

Today’s systems will create tremendous value in the world and, while they do have risks, the level of those risks feel commensurate with other Internet technologies and society’s likely approaches seem appropriate.

By contrast, the systems we are concerned about will have power beyond any technology yet created, and we should be careful not to water down the focus on them by applying similar standards to technology far below this bar.

Public input and potential

But the governance of the most powerful systems, as well as decisions regarding their deployment, must have strong public oversight. We believe people around the world should democratically decide on the bounds and defaults for AI systems. We don't yet know how to design such a mechanism, but we plan to experiment with its development. We continue to think that, within these wide bounds, individual users should have a lot of control over how the AI they use behaves.

Given the risks and difficulties, it’s worth considering why we are building this technology at all.

At OpenAI, we have two fundamental reasons. First, we believe it’s going to lead to a much better world than what we can imagine today (we are already seeing early examples of this in areas like education, creative work, and personal productivity). The world faces a lot of problems that we will need much more help to solve; this technology can improve our societies, and the creative ability of everyone to use these new tools is certain to astonish us. The economic growth and increase in quality of life will be astonishing.

Second, we believe it would be unintuitively risky and difficult to stop the creation of superintelligence. Because the upsides are so tremendous, the cost to build it decreases each year, the number of actors building it is rapidly increasing, and it’s inherently part of the technological path we are on, stopping it would require something like a global surveillance regime, and even that isn’t guaranteed to work. So we have to get it right.